www.ijarp.com ISSN 2456-9992 Page: 01-15

Research Article

International Journal Advanced Research Publications

Volume: 02

Issue: 10

ANALYSIS AND PREDICTION OF SALES, PROFITABILITY AND PRODUCT RETURNS IN E-COMMERCE USING MACHINE LEARNING

[1]Dr. G. Ramesh, [2]B. P. Adhithyan, *[3]S. Ramprakash

[1]Professor, [2,3]Undergraduate Scholars

[1-3]Department of Information Technology

[1-3]K.L.N. College of Engineering, Pottapalayam, Sivaganga, Tamil Nadu, India.

Article Received: 09 September 2026, Article Revised: 29 September 2026, Published on: 19 October 2026

***Corresponding Author: S. Ramprakash**

Undergraduate Scholars, Department of Information Technology K.L.N. College of Engineering, Pottapalayam, Sivaganga, Tamil Nadu, India.

Doi: https://doi-doi.org/101555/ijarp.4362

2. ABSTRACT

E-commerce companies collect large amounts of order data, but using it to make sound decisions on sales planning, discounts and returns remains difficult. This study analyses 34,500 orders placed between September 2023 and September 2025 across seven product categories and five regions, using data analysis and machine learning. Total net sales were 5.87 million after discounts of 308,518 (5.0% of sales). Electronics contributed 56.6% of sales, while Grocery lost money on average, with 95.5% of its orders unprofitable. The top third of customers generated 70.7% of revenue. Discounts were associated with lower profit per order but not with more returns. A LightGBM model predicted order-level net sales with $R^2 = 0.987$ and MAE = 4.76, but this is largely because net sales are calculated directly from price, quantity and discount, so the model mainly reproduced a known formula. A second LightGBM model, using lagged sales features, forecast monthly net sales between 233,000 and 262,000 for September 2025 to February 2026. Random Forest and Extra Trees models predicting whether an order would be returned reached only about 0.60 ROC-AUC, indicating that the available attributes carry little information about returns. The study also discusses its limitations and presents a Streamlit application and a Power BI dashboard for sharing the results.

3. KEYWORDS: E-commerce analytics; Sales forecasting; LightGBM; Return prediction; Customer segmentation; Discount analysis.

4. INTRODUCTION

Online retail now creates a flow of order-level records that explain what is bought, by whom, at what price and discount, and whether the purchase is eventually returned. Managers rely on these records for three choices: how much to plan for in future sales, which products and discount levels actually earn money, and which orders or customers deserve extra attention because of return risk. Gradient-boosted tree ensembles such as LightGBM and XGBoost have become a popular choice for tabular prediction problems of this kind [1, 2], but a prediction model is only useful when it is built on a clean data pipeline, evaluated honestly and connected to descriptive analysis that explains the business context.

This paper presents an end-to-end analytics study on a data set of 34,500 e-commerce orders placed between 12 September 2023 and 11 September 2025. The project covers data collection and quality checking, preprocessing and feature engineering, storage in a PostgreSQL database, exploratory sales analysis, supervised learning for order-level sales, a lag-feature model for monthly sales forecasting, a classifier for return risk and a set of business analyses on discounts, profitability and customer value. The outputs are delivered through a Streamlit application and a Power BI dashboard.

The objectives of the study are: (i) to describe sales, discount, profit and return behaviour by product category, region, payment method and customer segment; (ii) to build and evaluate a model for net sales; (iii) to produce a six-month forecast of monthly net sales; (iv) to assess whether order attributes available at purchase time can predict returns; and (v) to state clearly what the evidence does and does not support. The last objective matters because two of the models discussed in Sections 6.7 and 6.9 produce results that must be read with care.

The main contributions are a module-by-module analytics pipeline; a quantified view of where this retailer's revenue and losses are concentrated; and a transparent evaluation that separates a high headline score ($R^2 = 0.987$ for net sales) from the weak signal found for return prediction (ROC-AUC ≈ 0.60). The remainder of the paper is organised as follows. Section 4 also gives background and related work, Section 5 describes the data and methods, Section 6 reports and discusses the results and limitations, and Section 7 concludes.

4.1 Background and Related Work

Tree-based learning on tabular data. Random forests average many decorrelated decision trees grown on bootstrap samples, and extremely randomised trees add random split

thresholds to reduce variance further. Gradient boosting builds trees sequentially so that each tree corrects the residual error of the ensemble. LightGBM makes this process faster with histogram-based split search, gradient-based one-side sampling and exclusive feature bundling, and is commonly used for regression on structured business data. These techniques are available in the scikit-learn and LightGBM libraries used here.

Forecasting. Retail sales series are usually trend-seasonal and short. Standard practice is to compare any model with benchmarks, to create lag and rolling-mean features from past values, and to test on a chronologically later hold-out period instead of a random split. Multi-step forecasts made recursively feed each prediction back as a lag input, so errors can accumulate over the horizon.

Product returns. Product returns are a cost in online retail, and the reasons include product category, price, customer experience and the retailer's own return policy. Return events are usually rare, so accuracy is a poor metric: a model that predicts "no return" for every order can score very highly. Precision-recall analysis is recommended for imbalanced binary problems, and the ROC curve summarises ranking quality across thresholds.

Data leakage. Leakage occurs when a model is given information that would not be available at prediction time, or when the target can be computed directly from the inputs, which inflates apparent accuracy. We return to this issue in Section 6.7 because it influences how the order-level sales model should be understood.

5. MATERIALS AND METHODS

5.1 Dataset

The raw file contains 34,500 order records with 17 attributes: order, customer and product identifiers; product category; unit price; discount rate; quantity; payment method; order date; delivery time in days; region; return status; total amount; shipping cost; profit margin; customer age; and customer gender. The data cover 7,903 customers and 24,912 distinct products in seven categories (Beauty, Electronics, Fashion, Grocery, Home, Sports and Toys) and five regions (Central, East, North, South and West). Orders span 12 September 2023 to 11 September 2025, so the first and last calendar months are incomplete. Quality checks found no missing values and no duplicate rows, and no negative prices, quantities or shipping costs. [Authors to state the origin, licence and collection method of the data set here.]

Profit margin is recorded as a monetary amount per order rather than a percentage. Monetary values are reported in the data set's currency units throughout this paper.

5.2 Preprocessing and Feature Engineering

Preprocessing was performed in Python with pandas. The order date was converted to a date type and decomposed into year, month, day of month, day of week and ISO week of year. The return flag was coded as 1 for returned orders and 0 otherwise, and the discount rate was also expressed as a percentage. Gross sales, net sales and profit status were derived as

$$Gross_i = P_i \times Q_i \; ; \quad Net_i = P_i \times Q_i \times (1 - d_i) \tag{1}$$

where P_i is the unit price, Q_i the quantity and d_i the discount rate of order i. An order was labelled *Profit* if its profit margin was greater than zero and *Loss* otherwise. The cleaned file has 27 columns and was loaded into a PostgreSQL table, from which the modelling scripts read data.

5.3 Business Analysis Modules

Five descriptive analyses were run on the cleaned data. (a) *Monthly sales growth* compares each month with the previous one:

$$g_t = (S_t - S_{t-1}) / S_{t-1} \times 100 \tag{2}$$

where S is monthly net sales. (b) *Discount impact* measures the share of sales given away, DI = (Gross − Net) / Gross × 100, by discount level and by category. (c) *Profit and loss* compares mean profit margin and the number of profit and loss orders by category and product. (d) *Return analysis* computes return rates by category, region, payment method and discount level, and the net sales value of returned orders. (e) *Customer value and segmentation* aggregates orders to customer level, defines customer value as net sales divided by number of orders, and splits customers into three equal-sized groups (Low, Medium and High Value) by total net sales using terciles.

5.4 Order-Level Net Sales Model

We trained a regressor to predict net sales using twelve features: price, discount, quantity, delivery time, shipping cost, customer age, year, month, day, day of week, week of year and discount percentage. The data were split randomly into 27,600 training and 6,900 test records (80/20, random seed 42). The LightGBM regressor used 200 trees, a learning rate of 0.05 and

a maximum depth of 7. Performance was measured with mean absolute error, root mean squared error and the coefficient of determination:

$$MAE = (1/n)\,\Sigma\,|y_i - \hat{y}_i|\;;\quad RMSE = \sqrt{[(1/n)\,\Sigma\,(y_i - \hat{y}_i)^2]} \tag{3}$$

$$R^2 = 1 - \Sigma\,(y_i - \hat{y}_i)^2 / \Sigma\,(y_i - \bar{y})^2 \tag{4}$$

5.5 Monthly Sales Forecasting Model

We aggregated sales by month and removed the month of September 2025, which left 24 months from September 2023 to August 2025. The features were year, month, a time index, net sales from the previous one, two and three months (lag 1, 2 and 3) and the average of the past three months. After the first three months were dropped because their lags are undefined, 21 monthly observations remained; the last six were held out for testing and the earlier 15 used for training. A small LightGBM regressor (100 trees, learning rate 0.05, depth 3, 7 leaves, minimum 3 samples per leaf) was fitted. We then refitted the regressor on all 21 observations and used it recursively to forecast the next six months, from September 2025 to February 2026.

5.6 Return-Risk Classification

We treated return risk as a classification of the returned field. We used attributes that were available at the time of order: price, discount, quantity, customer age, order month, order day of week and one-hot encoded category, payment method, region and customer gender. Delivery time and shipping cost were excluded because they are not known at order time. Random Forest [3] and Extra Trees [4] classifiers with 200 trees, balanced class weights and a minimum of 5 samples per leaf were trained on a stratified 80/20 split (random seed 42); a further stratified split of the training portion was used as a validation set for choosing the decision threshold. Because returns are rare, the evaluation reports precision, recall and F1 for the returned class, together with the area under the precision-recall curve (PR-AUC) and the ROC curve (ROC-AUC) [8, 9]:

$$Precision = TP / (TP + FP)\;;\quad Recall = TP / (TP + FN) \tag{5}$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall) \tag{6}$$

where TP, FP and FN are the numbers of true positives, false positives and false negatives for the returned class.

6. RESULTS AND DISCUSSION

6.1 Overview of the Data

Table 1. Descriptive statistics of the numeric variables. (n = 34,500)

Variable	Mean	Std. dev.	Minimum	Median	Maximum
Unit price	119.39	195.62	1.01	45.66	2,930.47
Discount rate	0.05	0.07	0.00	0.00	0.30
Quantity	1.5	0.9	1.0	1.0	5.0
Delivery time (days)	4.8	1.2	3.0	5.0	13.0
Shipping cost	6.15	2.39	0.00	6.09	15.65
Customer age	43.5	15.0	18.0	43.0	69.0
Net sales per order	170.01	357.50	0.82	56.82	12,931.80
Profit margin per order	28.12	53.35	-6.20	10.55	1,536.17

Table 1 shows that order values are strongly right-skewed. The mean net sales per order (170.01) is three times larger than the median (56.82), and the largest order reached 12,932. Most orders do not have a discount; 54.9% of orders have no discount, and the average discount is 4.9%. Gross sales reached 6,173,811 while net sales were 5,865,293. This means discounts cut 308,518, or 5.0%, from sales. In total, 1,903 orders (5.52% of all orders) were returned and 6,122 orders (17.7%) were loss-making. Total recorded profit was 970,019, which means each order earned an average of 28.12.

6.2 Sales Trend and Category Mix

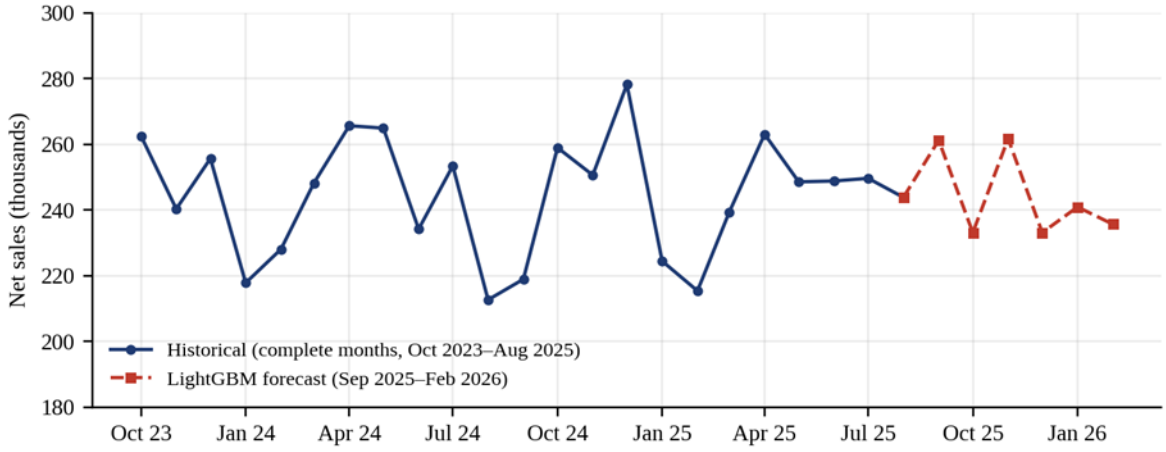


Fig. 1. Monthly net sales for complete months and the six-month LightGBM forecast

Monthly net sales for the months from October 2023 to August 2025 average 244,448 with a standard deviation of 18,151. The lowest month in this period is August 2024 with sales of

212,613. The highest month is December 2024 with sales of 278,154. Total net sales for the eleven months from October to August were 2,682,973 in 2023–24 and 2,720,462 in 2024–25, a change of +1.4%, so the business is essentially flat over the period. The month-on-month growth series shows volatility, with values ranging from −19.3% in January 2025 to [value]% in October 2024, after excluding two figures that are artefacts of partial months. These two figures in the growth table should not be interpreted as business events. The +73.7% increase in October 2023 and the −62.3% fall in September 2025 are due to those months being only partially covered by the available data, which begin on 12 September 2023 and end on 11 September 2025.

Table 2. Sales, returns and profitability by product category.

Category	Orders	Net sales	Share (%)	Return rate (%)	Mean margin	Loss orders (%)
Electronics	6,180	3,319,207	56.6	7.30	55.72	1.1
Home	5,487	1,077,681	18.4	5.65	47.86	3.2
Sports	4,171	629,825	10.7	4.94	38.49	2.5
Fashion	6,254	471,546	8.0	8.28	20.60	8.2
Beauty	4,103	153,019	2.6	3.78	11.99	11.7
Toys	4,247	132,014	2.3	4.94	7.93	21.2
Grocery	4,058	82,000	1.4	1.31	-2.26	95.5
All categories	**34,500**	**5,865,293**	**100.0**	**5.52**	**28.12**	**17.7**

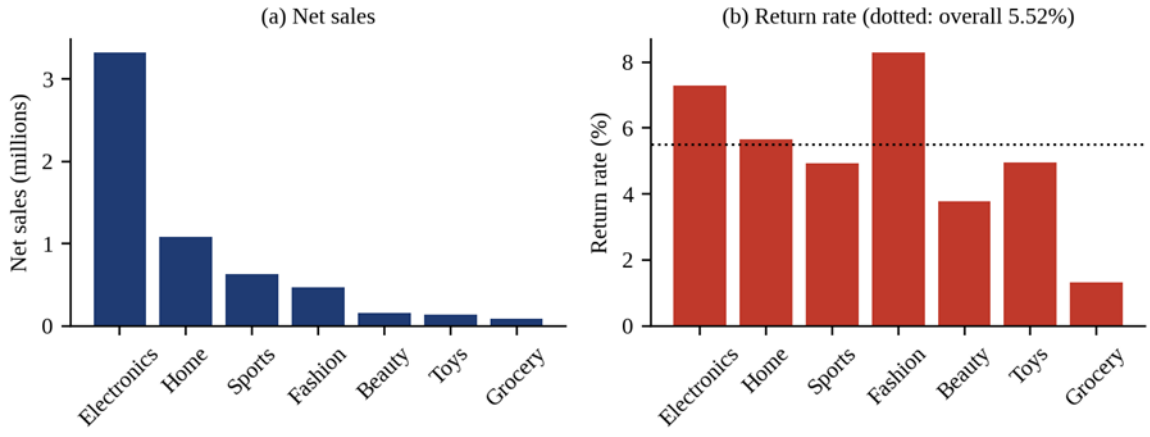


Fig. 2. Net sales and return rate by product category.

Revenue is highly concentrated by category (Table 2, Fig. 2). Electronics accounts for 56.6% of net sales from 17.9% of orders, and Electronics and Home together provide 75.0%.

Grocery, Toys and Beauty together contribute only 6.3% of net sales while accounting for 36.0% of orders. By region, net sales are fairly even: South (22.1%), North (21.6%), West (20.2%), East (20.1%) and Central (16.0%).

6.3 Discount Impact

Table 3. Orders, discount cost, profit and return rate by discount level.

Discount	Orders	Share (%)	Discount given	Mean margin	Return rate (%)
0%	18,939	54.9	0	30.07	5.41
5%	6,176	17.9	53,790	27.46	6.02
10%	4,159	12.1	74,685	26.04	5.46
15%	2,795	8.1	77,155	24.95	5.72
20%	1,727	5.0	62,901	22.67	5.15
30%	704	2.0	39,986	19.43	4.26

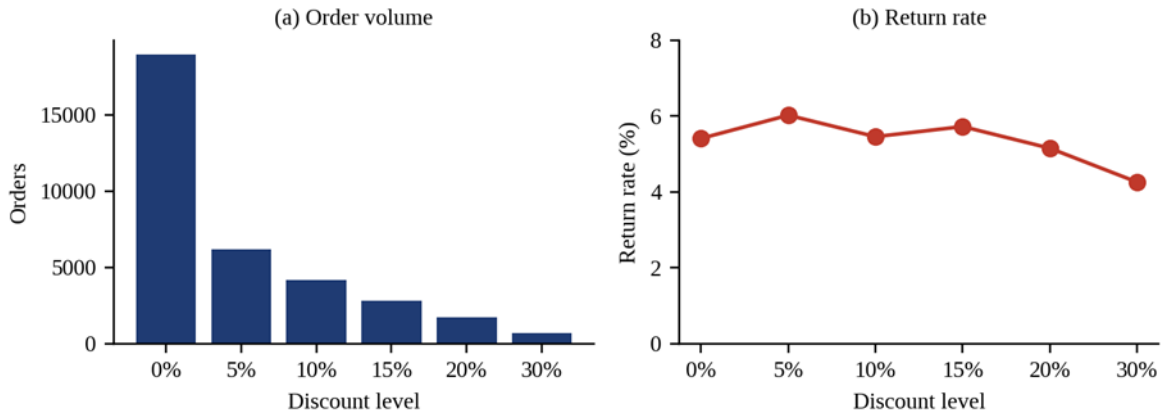


Fig. 3. Order volume and return rate by discount level.

Discounts of 5% to 30% were applied to 45.1% of orders and cost 308,518 in total (Table 3). The cost is spread across levels: the 10% and 15% tiers were the most expensive in absolute terms (74,685 and 77,155). Mean profit per order decreases as the discount increases. With no discount the average profit per order was 30.07; at a 30% discount it dropped to 19.43. This shows an association, not proof of cause. It is possible that deeper discounts are used on lower-priced items or items with lower margins, which could affect the results. Return rates do not rise with the discount (5.41% at no discount and 4.26% at 30%), so there is no sign in this data that discounted orders are returned more often. The category-level discount impact is almost uniform at 4.7–5.2% of gross sales, which suggests discount policy is applied evenly across categories rather than targeted.

6.4 Profitability and Loss-Making Orders

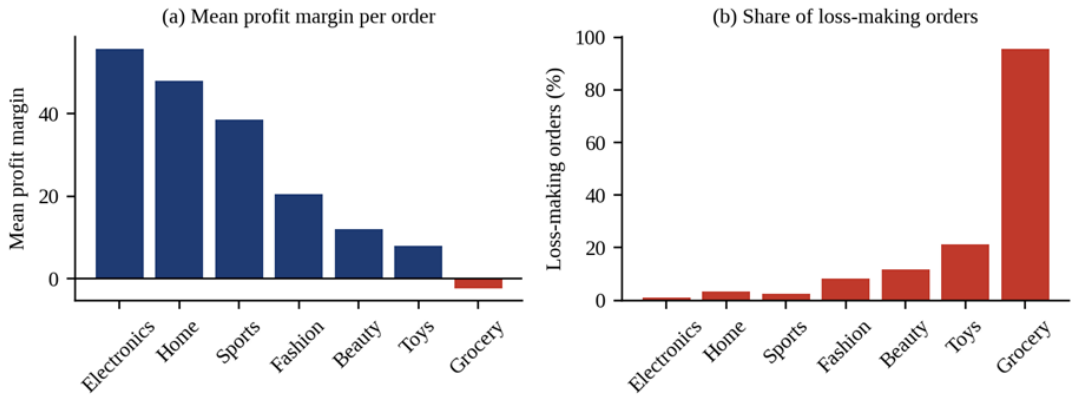


Fig. 4. Mean profit margin and share of loss-making orders by category.

Profitability differs sharply by category (Table 2, Fig. 4). Electronics earns the highest mean margin per order (55.72) and only 1.1% of its orders make a loss, followed by Home (47.86) and Sports (38.49). Grocery is the only category with a negative mean margin (-2.26), and 95.5% of its orders are loss-making. Toys is also weak, with 21.2% of orders at a loss. Because Grocery orders are small (mean net sales of about 20.21), fixed per-order costs such as shipping, which averages 6.15 per order, appear to absorb the margin. The data do not show the cost structure, so this idea is just a hypothesis that needs to be checked, not a proven fact. Profitability may be improved by reviewing pricing, basket sizes or shipping fees for Grocery and Toys.

6.5 Returns

Return rates range from 1.31% (Grocery) to 8.28% (Fashion) across categories, whereas the spread across payment methods (5.10% for cash on delivery to 5.78% for PayPal) and regions (5.10% in Central to 5.91% in East) is narrow. Delivery time is almost identical for returned and kept orders (4.84 versus 4.81 days). Returned orders represent 6.6% of net sales by value (388,756). Fashion has the highest return *rate*, but Electronics accounts for 63.0% of the *value* of returned sales because of its high prices, so value at risk is concentrated in a different place from return frequency. Returned orders had a slightly higher mean profit margin (32.48) than kept orders (27.86). The data do not say whether the recorded margin is reversed when an order is returned, so this should not be taken as returns being harmless.

6.6 Customer Segmentation

Table 4. Customer segments defined by terciles of total net sales.

Segment	Customers	Total net sales range	Orders	Net sales	Share (%)
Low Value	2,635	1.24 – 292.20	7,851	385,093	6.6
Medium Value	2,634	292.50 – 765.56	11,849	1,332,948	22.7
High Value	2,634	765.71 – 13,885.10	14,800	4,147,251	70.7
All customers	**7,903**	**–**	**34,500**	**5,865,293**	**100.0**

The top third of customers (High Value) generates 70.7% of net sales, whereas the bottom third (Low Value) generates 6.6% (Table 4). High Value customers also place the most orders per customer (5.6 against 3.0 for Low Value). Because the three groups have equal numbers of customers, the boundaries between them depend on this particular customer base. The heavy concentration of revenue in the High Value customers is a reason to act: protecting and serving them will keep most of the revenue, while Low Value customers are good targets for basket-building offers.

6.7 Order-Level Net Sales Model

Table 5. LightGBM net sales model: eight most important features. (split counts)

Feature	Importance
price	1,840
quantity	1,197
shipping_cost	794
discount	737
day	195
customer_age	192
week_of_year	130
day_of_week	66

On the 6,900 test orders the LightGBM model reached MAE = 4.76, RMSE = 39.72 and R^2 = 0.9867. The mean net sales per test order is 168.50, so the MAE is 2.8% of the mean. The largest single error was 2,529.29 (an order with net sales of 3,825.50 predicted as 6,354.79) and the ten largest errors account for about 85% of the total squared error, which is why the RMSE is much larger than the MAE. Price, quantity, shipping cost and discount are the four most important features (Table 5); the date features contribute little and the duplicated discount percentage receives no splits.

These results must be interpreted carefully. Net sales are *defined* as price × quantity × (1 − discount) (Equation 1), and price, quantity and discount are all inputs to the model. In the cleaned data this identity holds to within rounding error for every order. The model is therefore learning to approximate a known arithmetic formula, not forecasting an uncertain outcome, and $R^2 = 0.987$ mostly measures how well trees approximate a multiplicative function. This is a form of target leakage. The model is useful as a calculator that reproduces order value from the order attributes. It confirms that price and quantity dominate order value, but it should not be presented as evidence of predictive power for future sales. A genuine order-level prediction task would exclude price, quantity and discount and would predict an outcome that is not an exact function of the inputs.

6.8 Monthly Forecast

Table 6. Six-month recursive forecast of monthly net sales.

Month	Forecast net sales
September 2025	261,189
October 2025	233,068
November 2025	261,665
December 2025	233,068
January 2026	240,871
February 2026	235,623

The forecast ranges from 233,068 to 261,665 per month with a six-month mean of 244,247 (Table 6, Fig. 1). This is very close to the historical average of 244,448. The model suggests a business environment with only small fluctuations and no growth, which matches the flat trend described in Section 6.2. October 2025 and December 2025 have identical forecast values (233,068). This happens because the model is a tree ensemble trained on just 21 data points and can only output a limited number of distinct levels. Three cautions apply. First, a series of 24 monthly values is too short for a learned model to separate trend, seasonality and noise, and the data cover only 23 complete months, so annual seasonality cannot be estimated reliably. Second, the recursive procedure feeds forecasts back as inputs, so uncertainty grows with the horizon, and no prediction interval is given. Third, the hold-out error of this forecasting model was computed by the project script but not saved with its outputs, and no naive benchmark (for example, repeating last month's value) was evaluated, so the forecast's accuracy relative to simple alternatives has not been established. These forecasts should be treated as planning scenarios, not commitments.

6.9 Return-Risk Prediction

Table 7. Return-risk classifiers on the held-out test set. (6,900 orders, 381 returned)

Model	Threshold	Accuracy	Precision	Recall	F1	PR-AUC	ROC-AUC
Random Forest	0.50	0.9028	0.0950	0.0892	0.0920	0.0751	0.6022
Extra Trees	0.50	0.8728	0.0824	0.1286	0.1004	0.0746	0.5980
Extra Trees	0.30	0.5297	0.0725	0.6378	0.1303	0.0746	0.5980

Precision, recall and F1 refer to the returned class. A classifier that always predicts "not returned" would have accuracy 0.9448 and a PR-AUC equal to the base rate, 0.0552.

Table 8. Confusion-matrix summary for the return-risk models.

Model	True not returned	False alerts	Missed returns	Detected returns
Random Forest (0.50)	6,195	324	347	34
Extra Trees (0.50)	5,973	546	332	49
Extra Trees (0.30)	3,412	3,107	138	243

The classifiers do only a little better than guessing (Table 7). At the 0.50 threshold both models have high accuracy (0.9028 and 0.8728), but this is misleading: 94.5% of test orders are not returned, so predicting no returns at all would give higher accuracy (0.9448) than either model. Their precision is 8 to 9.5% and their recall is 9 to 13%, which means they find about one in ten returns but also raise many wrong alerts. ROC-AUC is 0.60 and PR-AUC is about 0.075 against a base rate of 0.055, so the ranking of orders by risk is only slightly better than random.

Lowering the threshold to 0.30 for Extra Trees, chosen using the validation split, raises recall to 63.8% (243 of 381 returns found), but precision falls to 7.3% and 3,107 non-returned orders are flagged (Table 8). The flagged group therefore contains returns at 7.3%, only 1.3 times the overall rate of 5.5%. Whether this is worthwhile depends on the cost of an intervention (for example, an extra quality check or a confirmation message) relative to the cost of a return, which the data do not provide. The grouped feature importance of the Extra Trees model is spread widely across order timing (0.204), customer attributes (0.167), category (0.148), region (0.146) and payment method (0.141), with price (0.070), discount (0.071) and quantity (0.054) lower. Among individual features, order day of week, order month and customer age rank highest. This diffuse pattern, with timing and age ranking above price, is what a model fitting noise typically produces, and is consistent with the low

AUC. The more informative result is descriptive: category is the one attribute with a clear relationship to return rate (Section 6.5).

The practical conclusion is that with these attributes return risk cannot be predicted well enough for automated decisions. Better signals would probably come from information not in this data set, such as product-level return history, customer return history, product ratings and descriptions, size or fit details for Fashion, and reason codes for returns.

6.10 Managerial Implications

The analyses support five conclusions. First, revenue depends heavily on Electronics and on the top third of customers, so supply, pricing and service for these groups deserve attention. Second, Grocery (and to an extent Toys) makes less money on most orders, so its pricing, minimum order value or shipping charges need to be checked. Third, discounts are associated with lower profit per order without any clear effect on returns, and because the data have no information on whether discounts increased sales, their commercial benefit cannot be judged from this study and should be tested with controlled experiments. Fourth, Fashion should be targeted for return reduction by how returns happen, and Electronics by how much each return is worth. Fifth, the monthly forecast shows sales near 230–260 thousand per month, which is helpful for planning as long as the uncertainty is understood. The Streamlit application and the Power BI dashboard give these views to non-technical users through pages for historical analytics, sales prediction, return-risk prediction, future forecast and dashboard exploration.

6.11 Limitations

The study has limitations. The data cover two years and one retailer, and their origin is not documented in the project files; results may not apply to other businesses or time periods. Orders are the unit of analysis, so repeat behaviour over time is only captured through customer aggregates. The order-level net sales model is affected by target leakage as explained above. The monthly forecast is based on 21 points, has no prediction interval and no benchmark comparison, and its hold-out error was not kept. The return-risk models use a stratified split and fixed hyperparameters; no cross-validation, calibration or hyperparameter search was done, and the threshold analysis was done on one data set. The profit margin variable is a recorded amount whose cost components are unknown, and the discount comparison is observational. Finally, the figures reported here come from this data set and split and do not guarantee performance elsewhere.

7. CONCLUSION

This paper looked at 34,500 e-commerce orders with descriptive analytics and three machine-learning models. Net sales of 5,865,293 are concentrated in Electronics (56.6%) and in the High Value customer tercile (70.7%); Grocery loses money on most orders; and discounts cost 5.0% of gross sales and lower profit per order without increasing returns. A LightGBM model reproduces order-level sales with $R^2 = 0.987$, but this result is mostly because net sales are a function of the model's inputs. A lag-based model forecasts stable sales of about 233–262 thousand per month through February 2026, subject to the cautions above. Return risk could not be predicted well from the available attributes (ROC-AUC ≈ 0.60).

7.1 Future Work

Future work should (i) remove price, quantity and discount from any order-level prediction target or choose a non-deterministic target; (ii) evaluate the forecast against seasonal-naive and exponential-smoothing benchmarks with rolling-origin validation and prediction intervals once more history is available; (iii) add customer return history, product-level features and return reasons to the return model, with cross-validation and probability calibration; (iv) estimate the causal effect of discounts through experiments; and (v) add cost data to explain Grocery and Toys losses.

8. ACKNOWLEDGEMENTS

The authors would like to thank the management and faculty members of the Department of Information Technology for their support and guidance throughout this project. The authors also thank everyone who provided assistance and encouragement during the development of the E-commerce Sales Forecasting and Analysis project. Finally, the authors acknowledge the publicly available dataset used for the analysis and the open-source tools that supported the implementation.

9. REFERENCES

1. Demand Forecasting of Online Retail Using Machine Learning, AIP Conference Proceedings, 2024. (Base Paper)
2. Machine Learning Approaches for E-Commerce Product Return Prediction, Journal of Cleaner Production, 2024.
3. Demand Forecasting and Price Optimization for Online Retail, 2021.
4. Machine Learning-Based E-Commerce Sales and Demand Prediction, 2026.

5. LightGBM Documentation and Gradient Boosting Methods, Microsoft LightGBM Documentation.
6. Python Pandas Documentation, Data Cleaning and Data Analysis.
7. SQL Documentation, Data Storage, Queries and Aggregation.
8. Power BI Documentation, Microsoft.